

LIRA: Adaptive Contention-Aware Thread Placement for Parallel Runtime Systems

Alexander Collins^{*}, Tim Harris[†], Murray Cole^{*}, Christian Fensch[‡]

^{*} University of Edinburgh

[†] Oracle Labs, UK

[‡] Heriot Watt University



THE UNIVERSITY of EDINBURGH
informatics

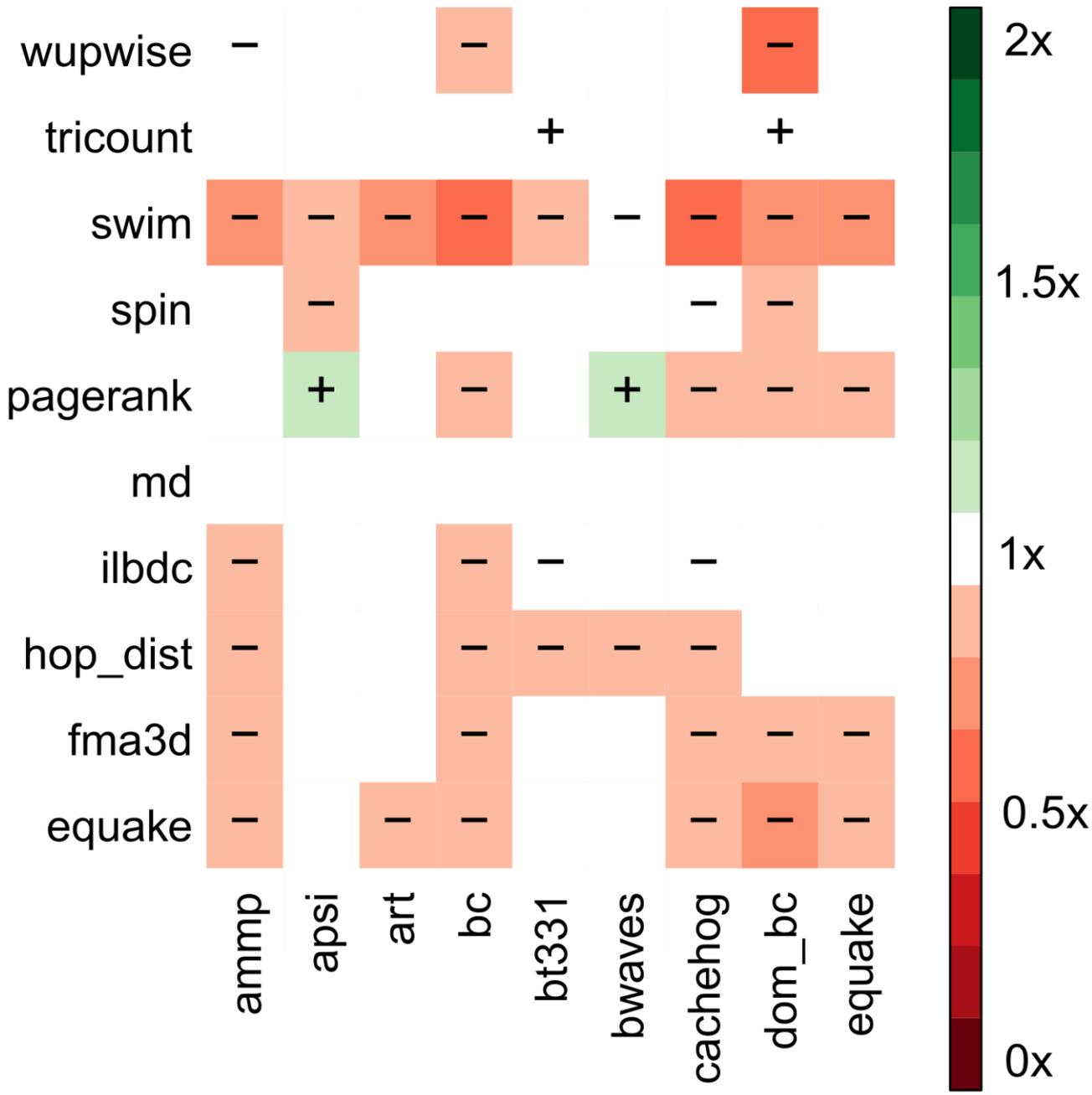
icsa

Institute for Computing
Systems Architecture

The Problem

- Multi-socket machines common-place
- Run multiple parallel programs
- Co-location affects performance
- ***Which programs should we co-locate?***

Program providing workload



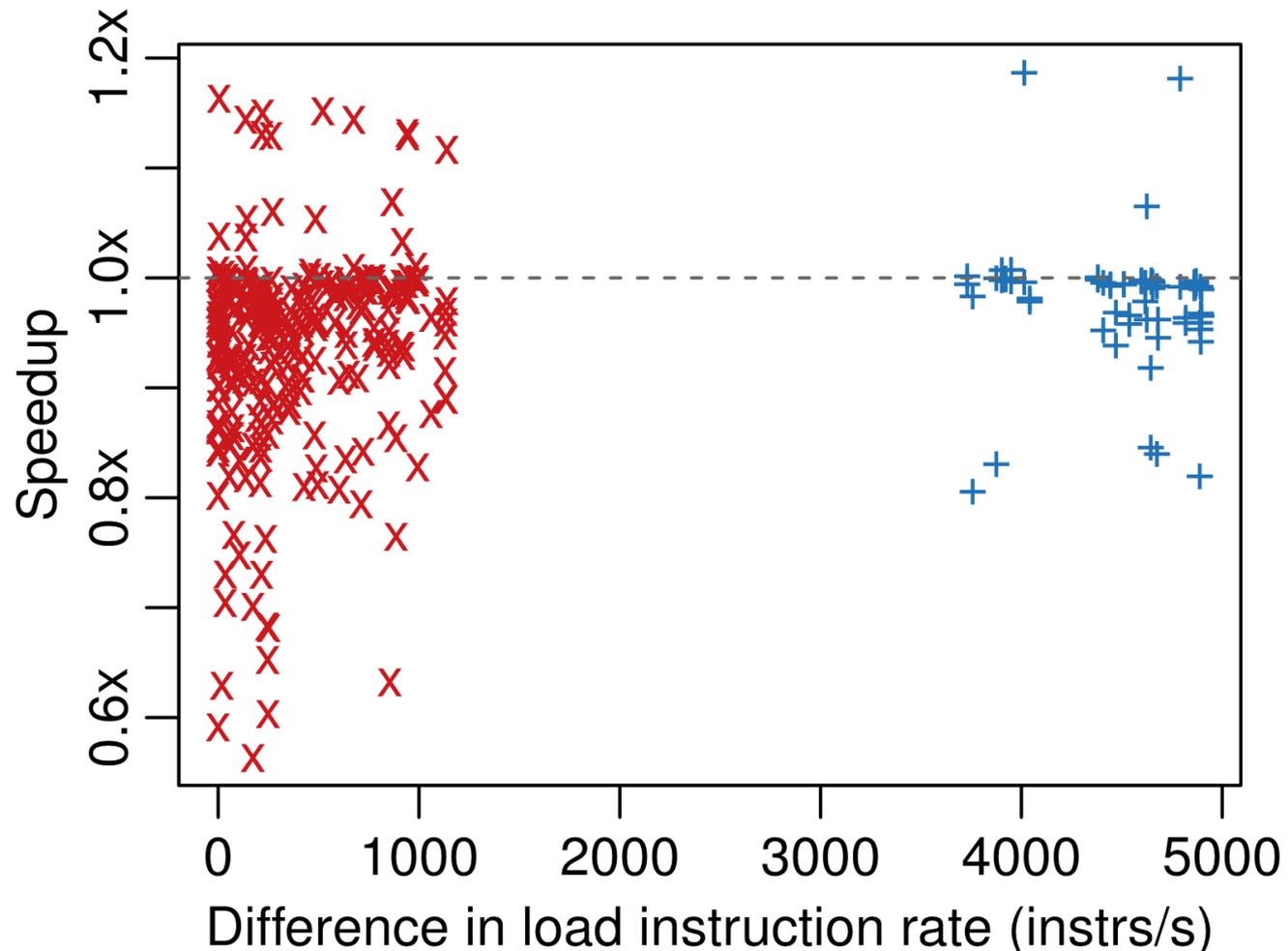
Program measured

The Problem

- System workload is constantly changing
- Best co-location changes
- ***Need an online adaptive solution***

Our Insight

- Balance load instruction rate across sockets



Our Solution

- Schedule programs to sockets
- Maximise difference in load instruction rate (LIRA heuristic)
- Built on top of Callisto^[1]
- Each program pins one thread to each core
- One thread on each core is high priority
- High priority thread runs unless it stalls

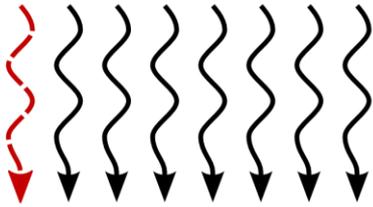
[1] Callisto: Co-scheduling Parallel Runtime Systems, Harris et al. EuroSys '14

Our Solution



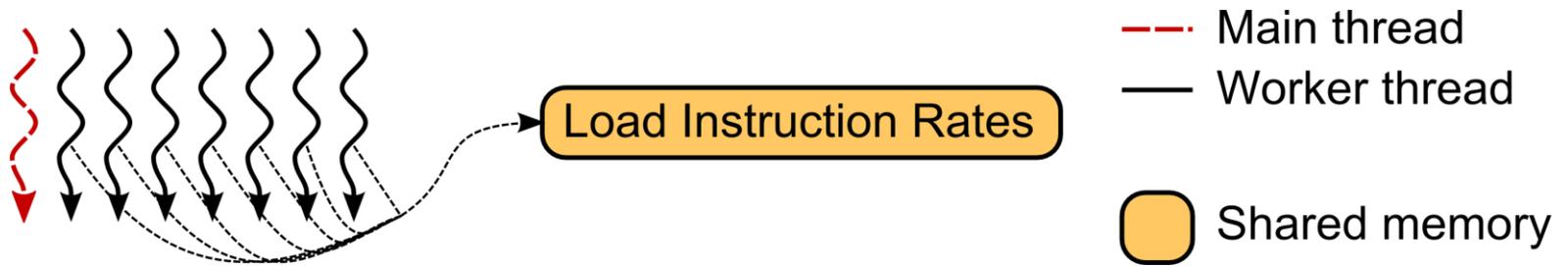
--- Main thread

Our Solution

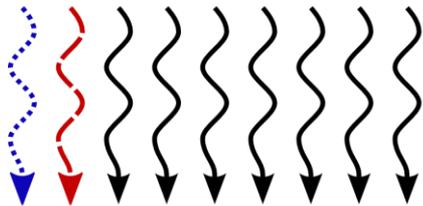


--- Main thread
— Worker thread

Our Solution



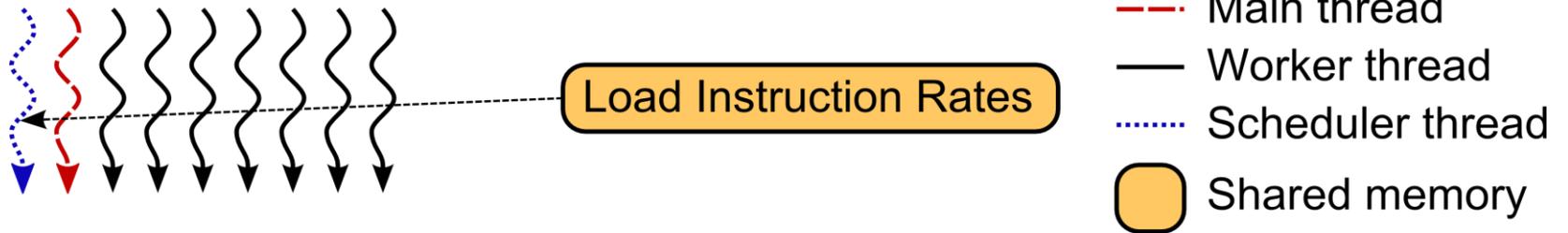
Our Solution



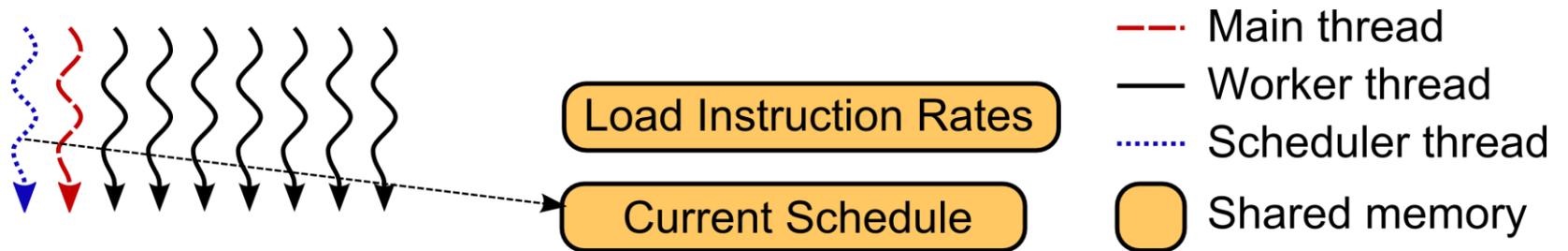
Load Instruction Rates

- Main thread
- Worker thread
- Scheduler thread
- Shared memory

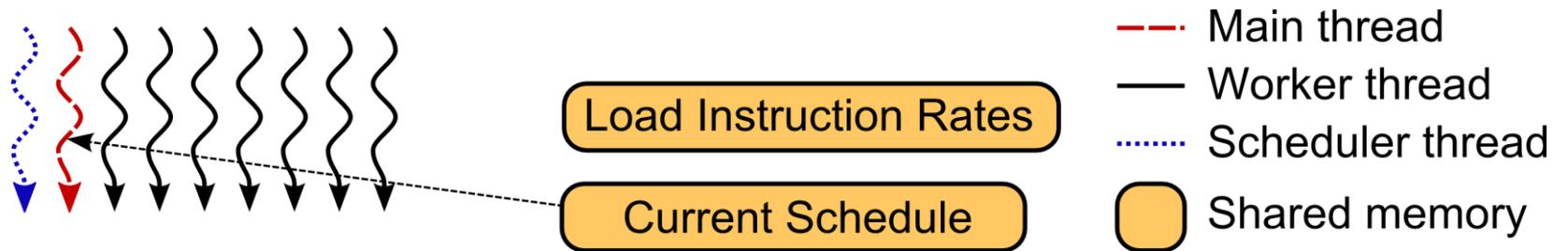
Our Solution



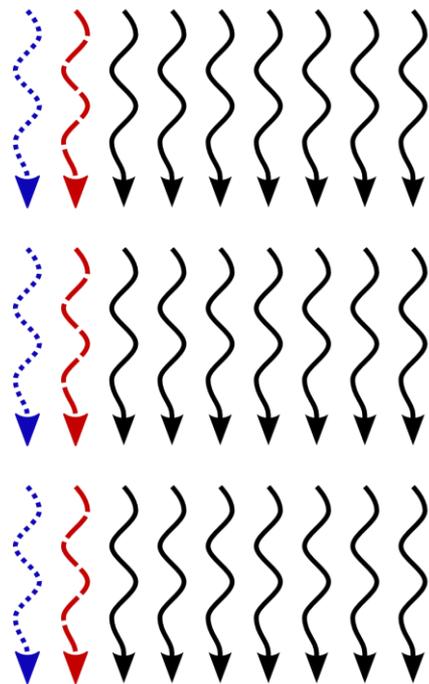
Our Solution



Our Solution



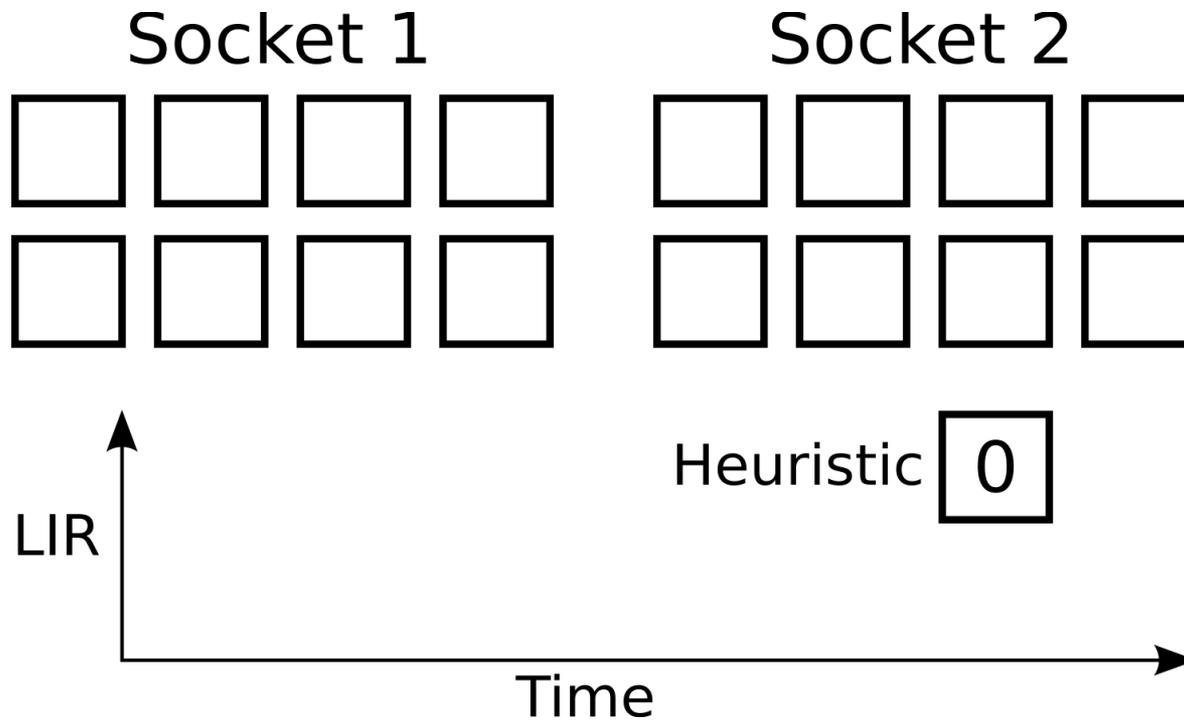
Our Solution

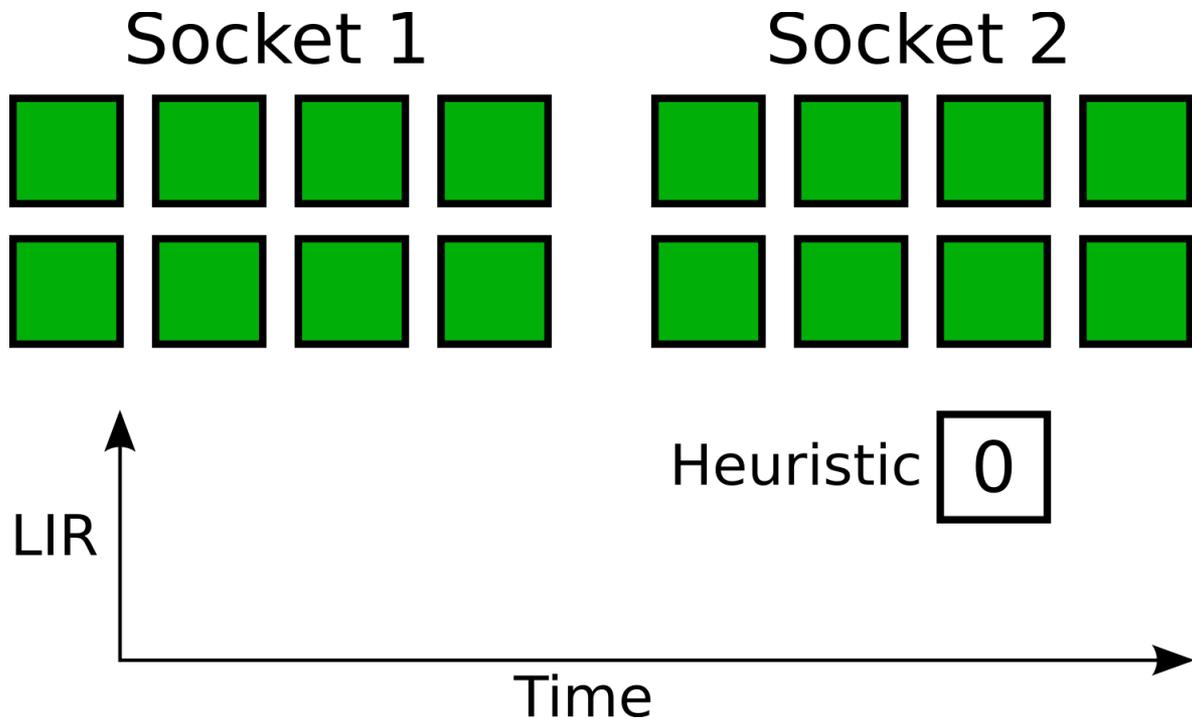


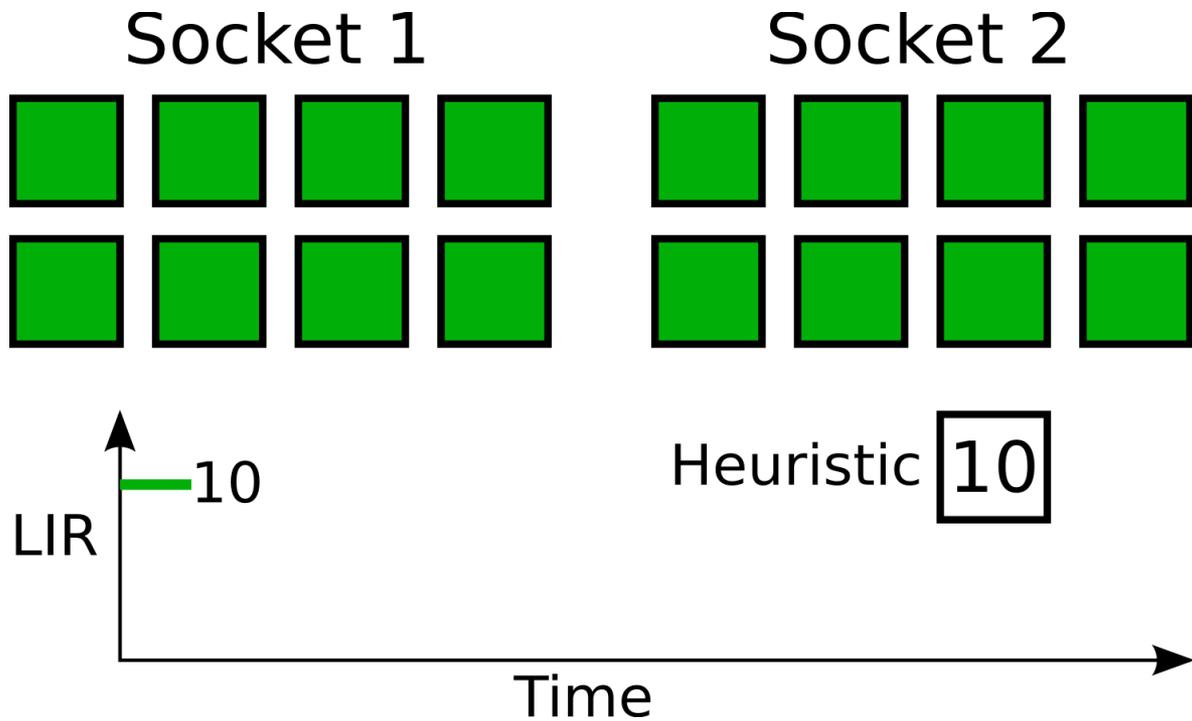
Load Instruction Rates

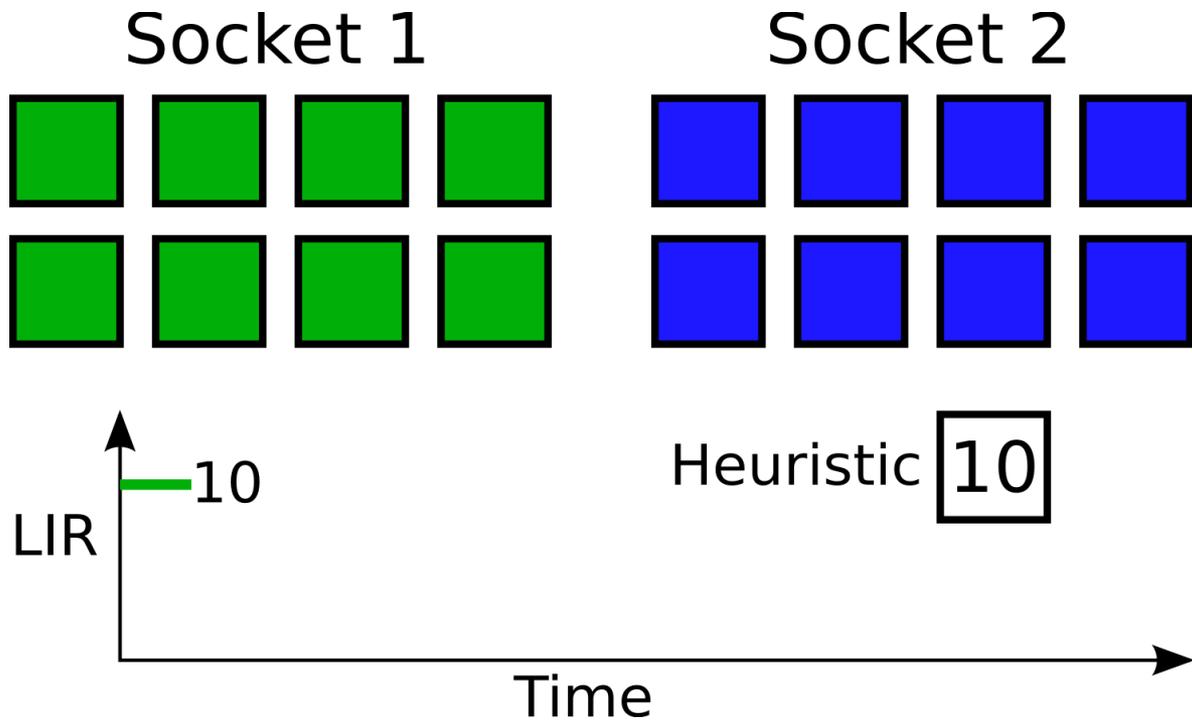
Current Schedule

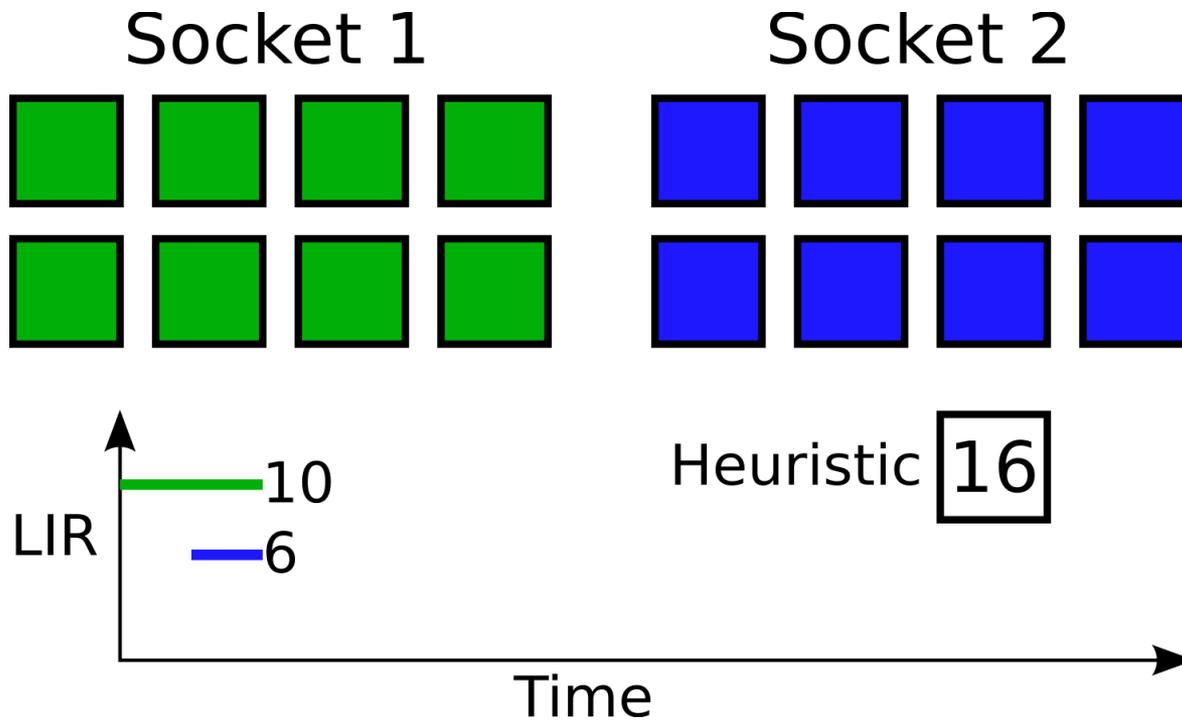
- - - Main thread
- Worker thread
- Scheduler thread
- Shared memory

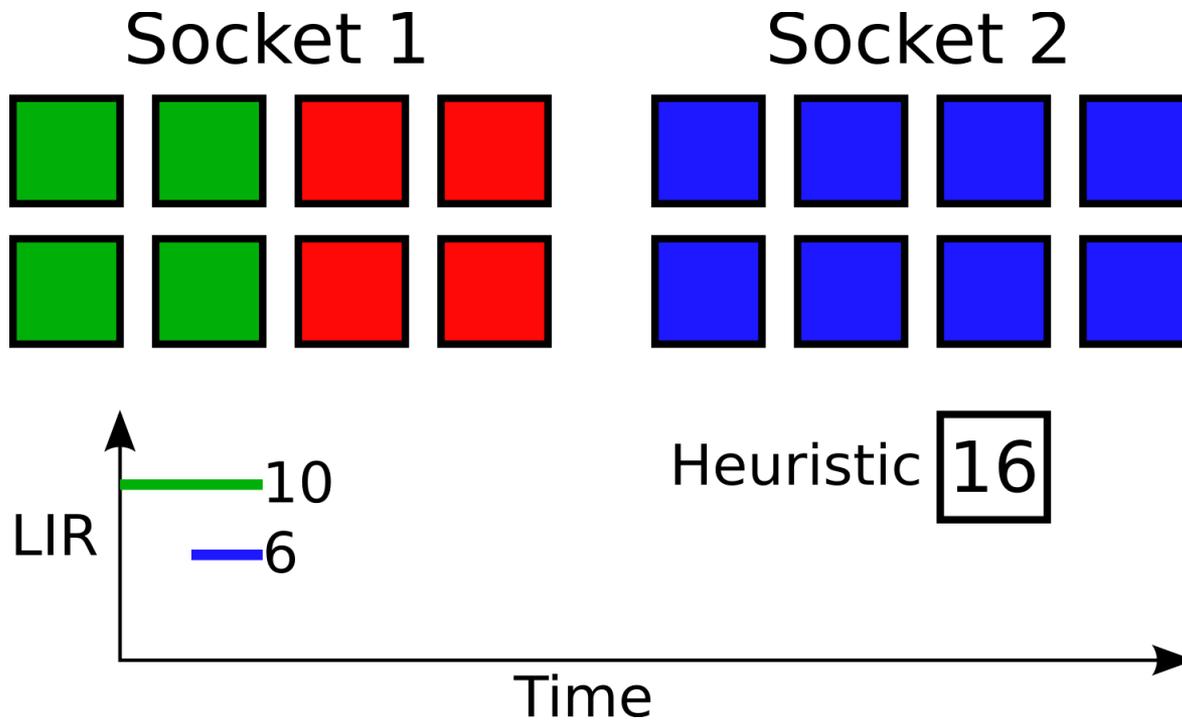


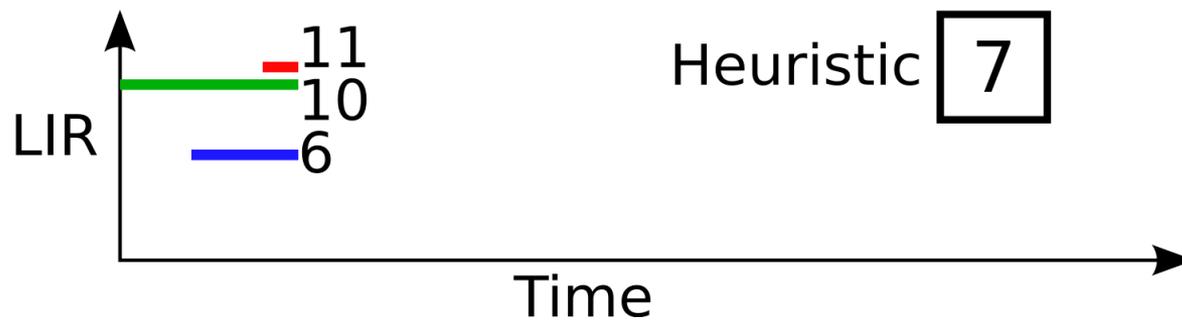
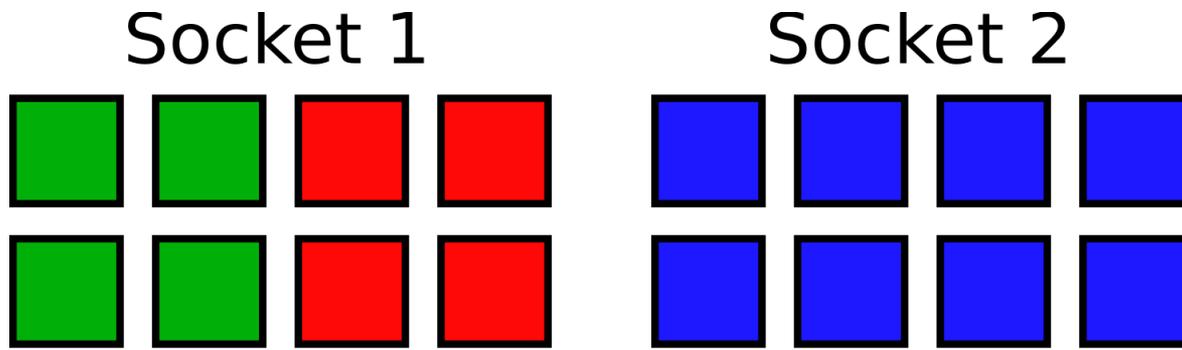


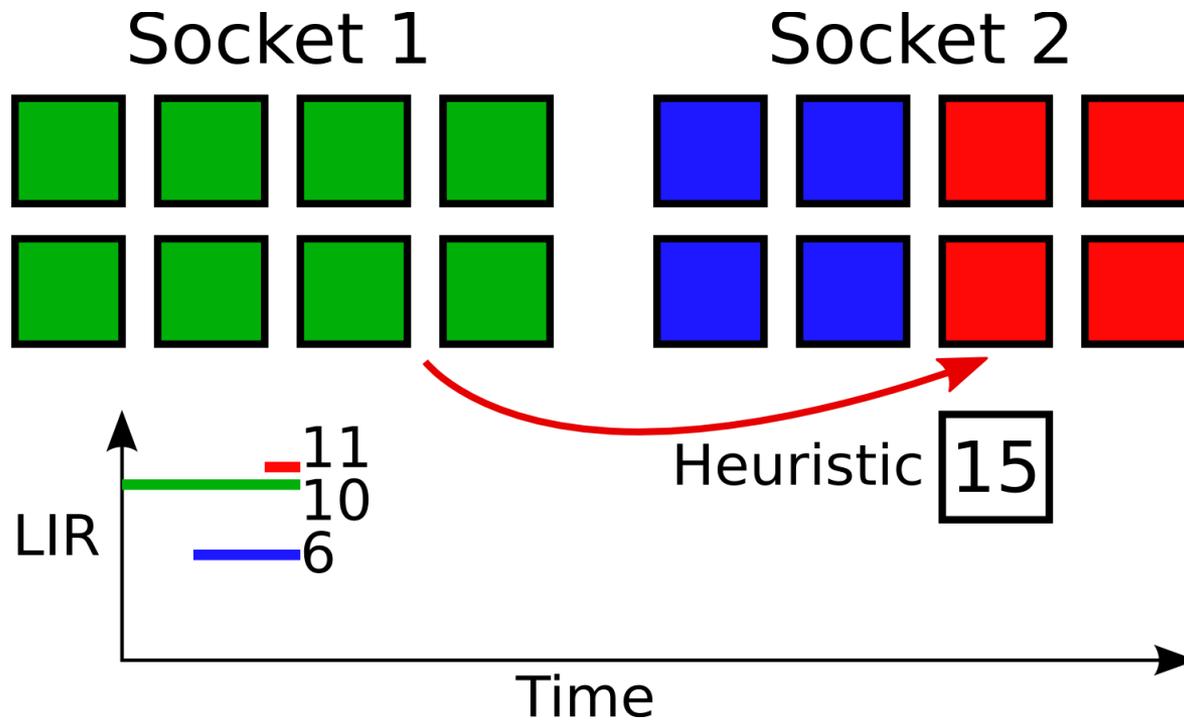




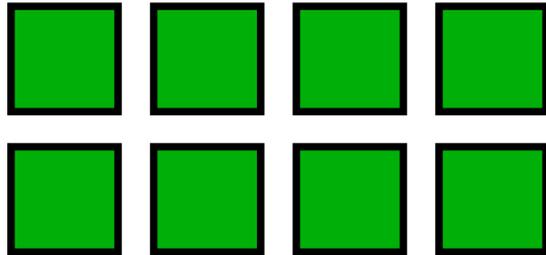




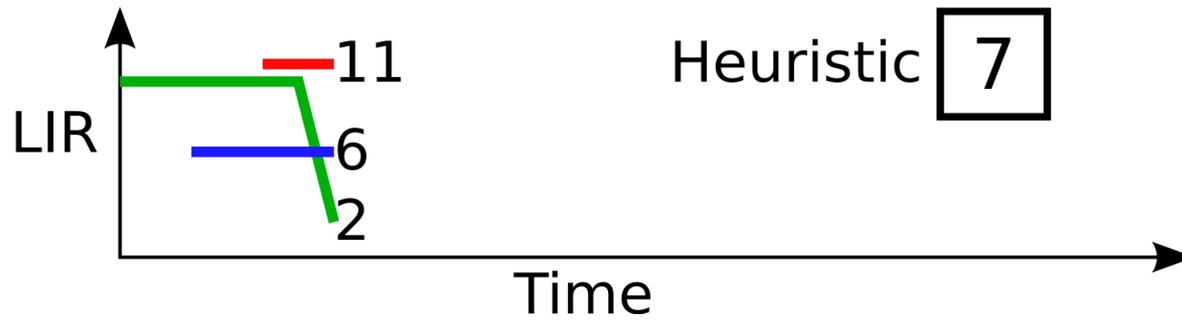
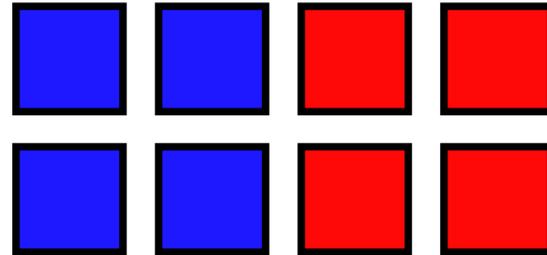


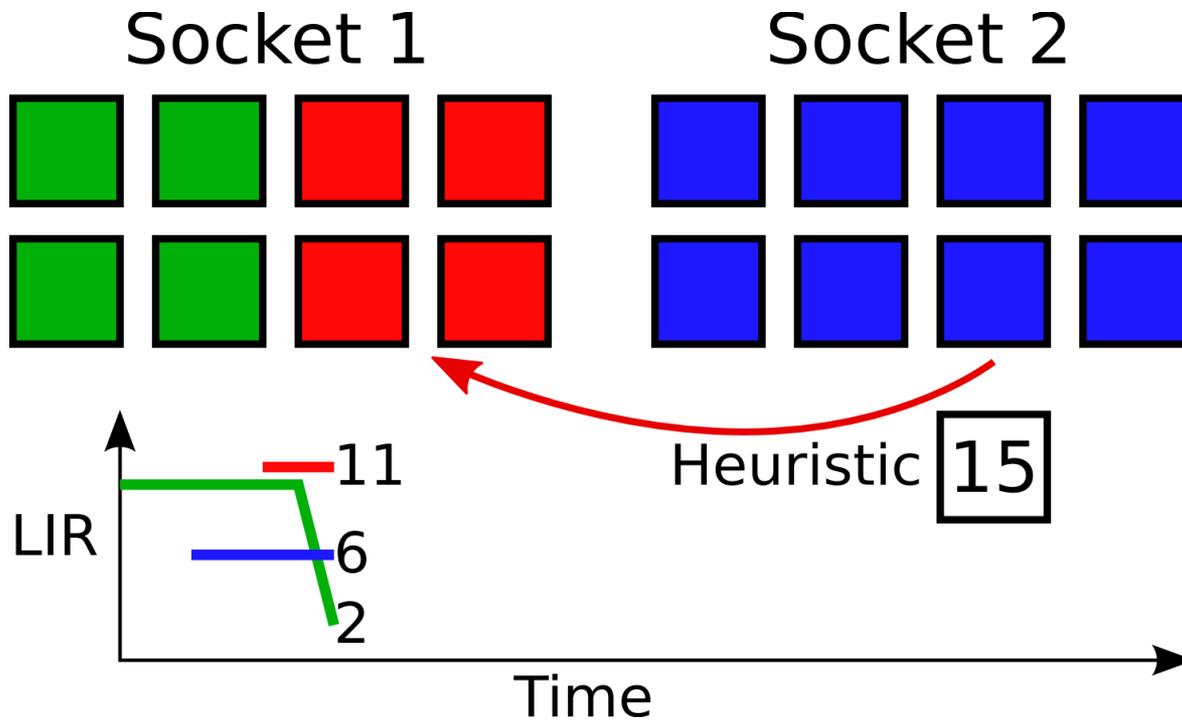


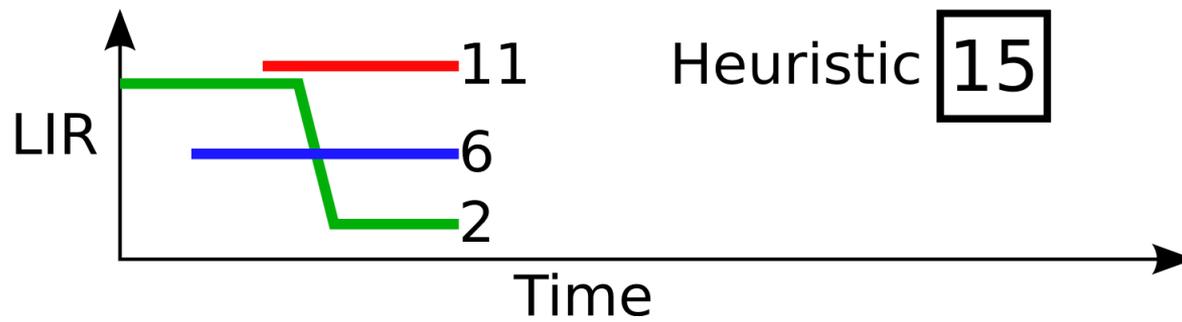
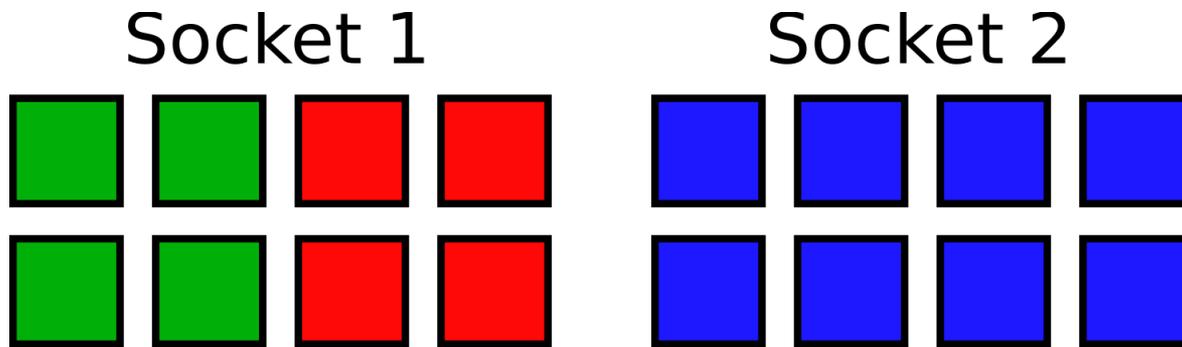
Socket 1

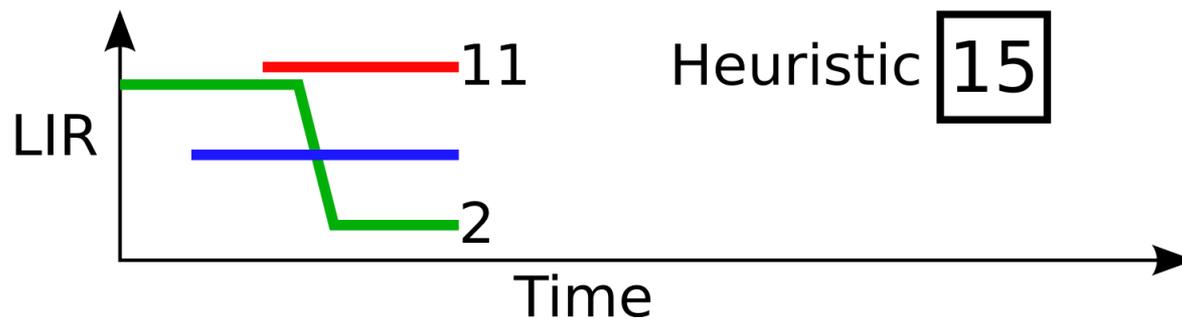
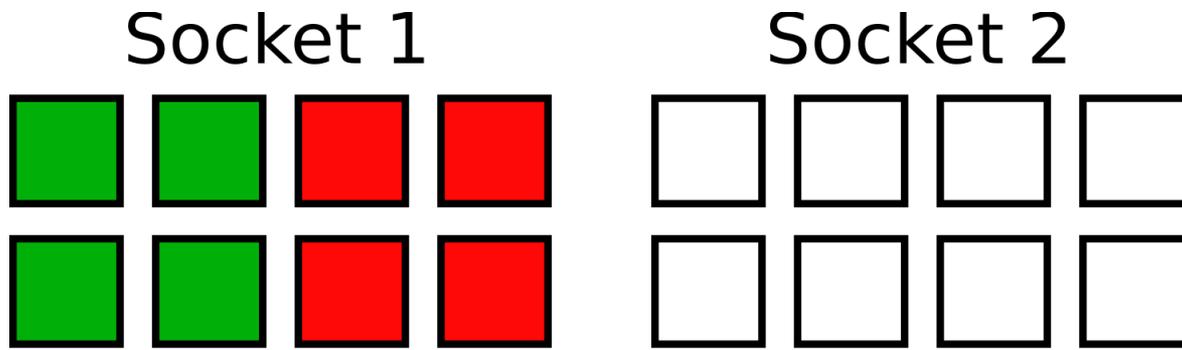


Socket 2

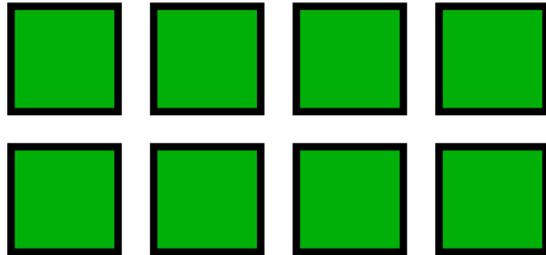




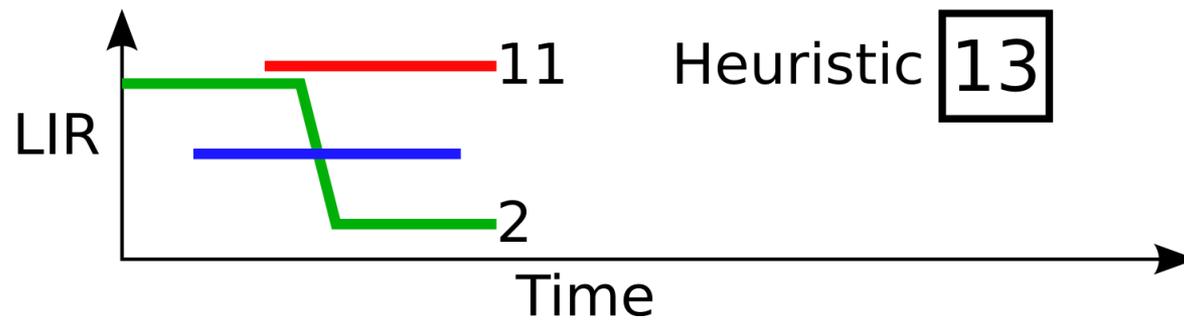
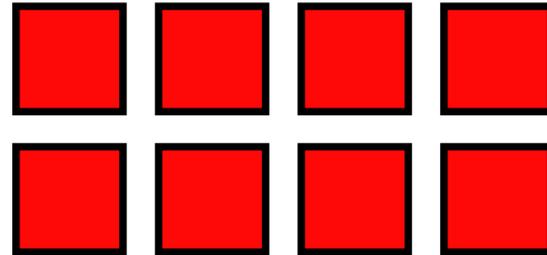




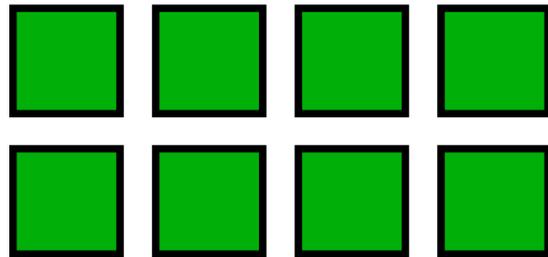
Socket 1



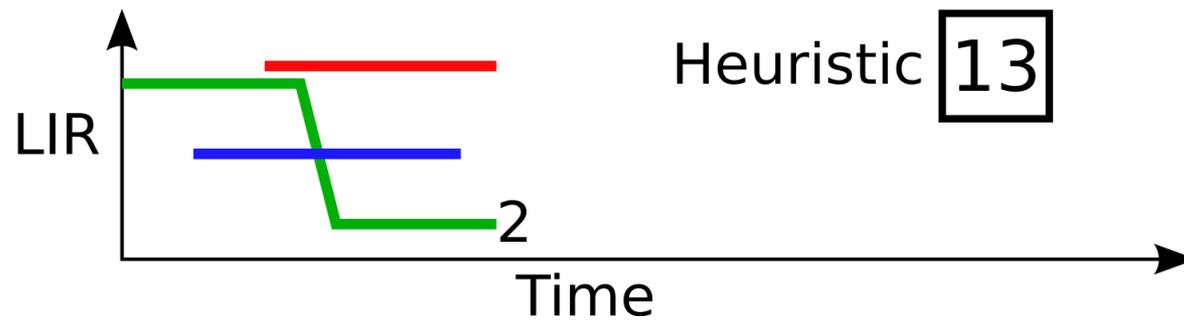
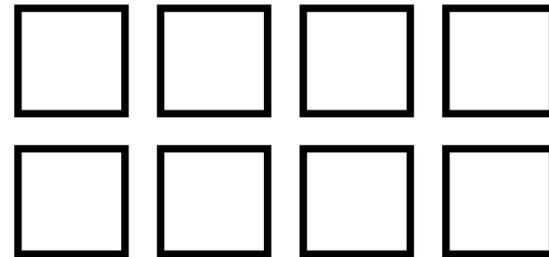
Socket 2

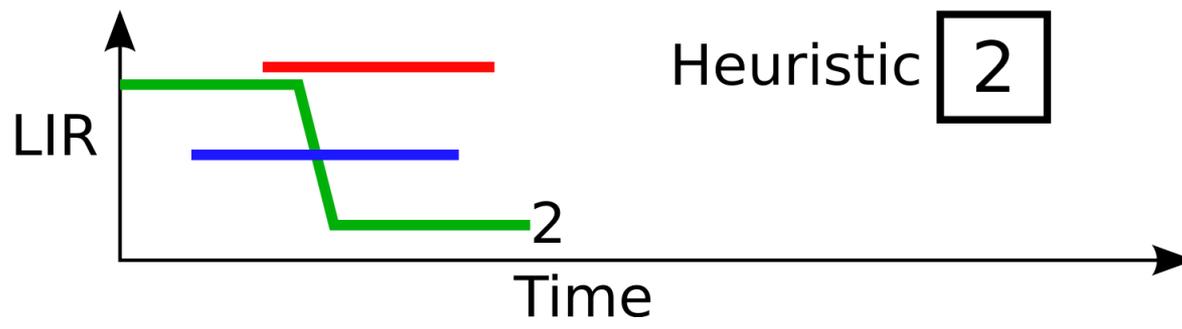
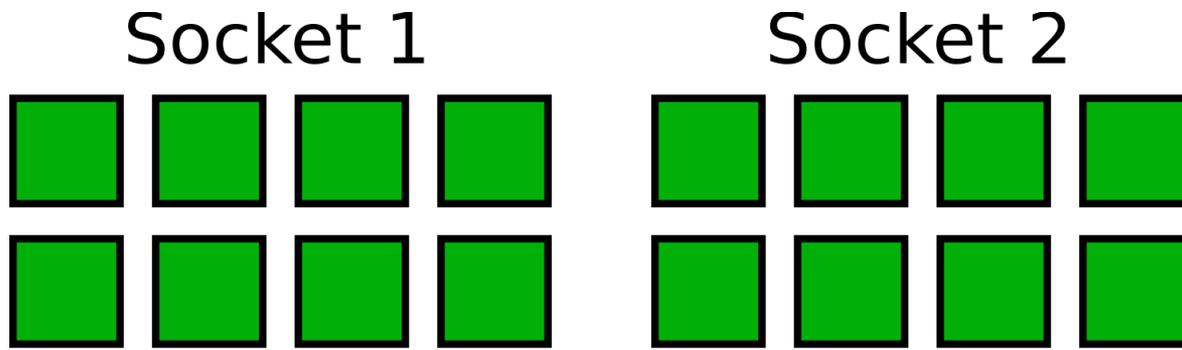


Socket 1



Socket 2





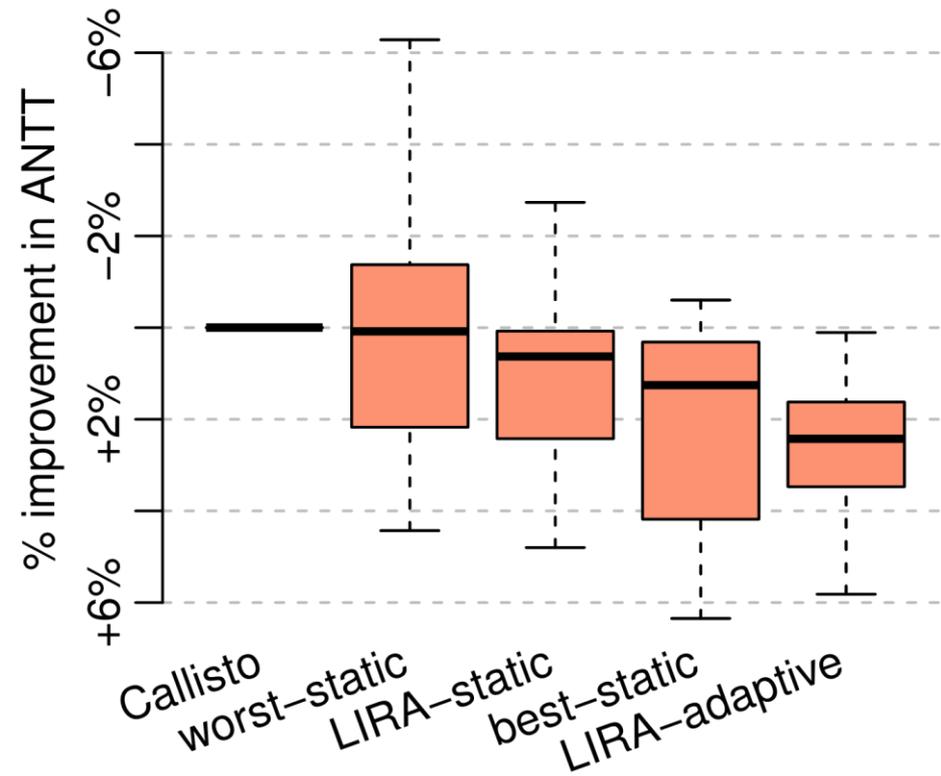
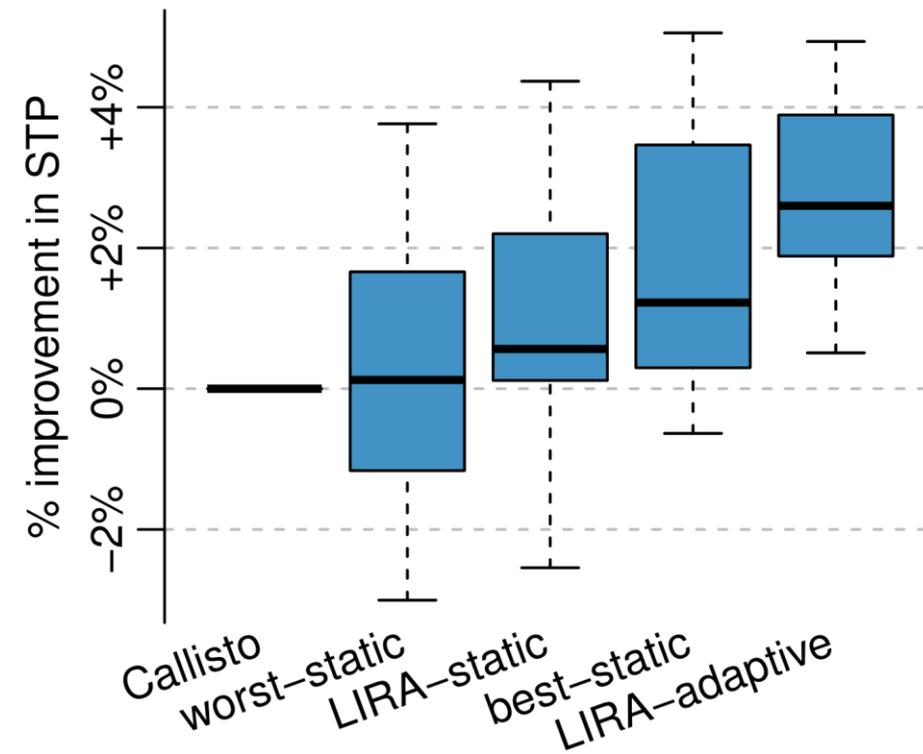
Evaluation

- 11 benchmarks from SPEC OpenMP 2001
- 4 from GreenMarl project
- 1 using CDDP (betweenness-centrality)
- Dual-socket Xeon E5-2660
- 8 cores each (hyperthreading disabled)

Evaluation

- Measure 32 combinations of four programs
- ANTT and STP system performance metrics
- Comparing:
 - Socket unaware Callisto
 - LIRA static tuning
 - LIRA adaptive tuning

Evaluation



Conclusions

- Co-location affects performance
- Adaptive online tuning is required
- LIRA heuristic improves performance

- More details in the paper

LIRA: Adaptive Contention-Aware Thread Placement for Parallel Runtime Systems

Alexander Collins^{*}, Tim Harris[†], Murray Cole^{*}, Christian Fensch[‡]

^{*} University of Edinburgh

[†] Oracle Labs, UK

[‡] Heriot Watt University



THE UNIVERSITY *of* EDINBURGH
informatics

icsa

Institute for Computing
Systems Architecture

